Verification to Determine and Measure Forecasting Skill

IRVING I. GRINGORTEN

Air Force Cambridge Research Laboratories, Bedford, Mass.

(Revised manuscript received 23 February 1967)

ABSTRACT

Three distinct purposes for the verification and scoring of forecasts have been generally recognized: determination of the accuracy of the forecasts, their operational value, and, lastly, the skill of the forecaster. Although the question most frequently asked is "How accurate?," the answer, "usually about 85%," is the most trivial. The operational value of forecasts is much more interesting and significant, but difficult or impossible to determine. The skill of the forecaster, however, is a tractable subject. It is defined as "The ability of the forecaster to sort or group the weather situations so that within any group the probability of one out of several mutually exclusive subsequent events is increased above its climatic frequency." A set of scores can be designed to reward the forecaster for skilled grouping or sorting of weather patterns and to permit no advantage to an unskilled strategy. Such a scoring system was described more than 10 years ago, but it is not popularly accepted, partly because there never has been a set of uniform goals for verification.

1. Introduction

In this paper, verification is limited to the forecast of several mutually exclusive events or categories. The subject is an old one, and several papers date back to the 19th century (Gringorten, 1951). But in the past 20 years three distinct purposes for verification have emerged. One purpose is to test the accuracy of forecasts. Another purpose is to test the operational or economic value of a set of forecasts. Still another purpose is to test the skill of the forecaster. These differences would vanish if forecasting could approach perfection. But perfect forecasts do not exist. This paper is devoted to the measure of skill of much, much-less-than-perfect forecasts, indeed, to determine if there exists any skill at all in a set of forecasts.

2. Purposes

To distinguish between the three purposes of verification let us examine a matrix of scores (Table 1). The element A_{kj} is the score for the forecast of the kth category X_k when the jth category X_j was ultimately observed. The rows, instead of being designated as forecast categories, could be considered as alternative courses of action. If the scores in the kth row are multiplied by the probabilities $P_0, \cdots, P_n$ of the n kinds of subsequent events and added,

TABLE 1. A matrix of scores, one row for each forecast and one column for each observed event of several mutually exclusive categories of events.

Forecast	Observed event			
event	X_0	X_1	X_2	X_3
X_0	A_{00}	A_{01}	A_{02}	A_{03}
X_1	A_{10}	A_{11}	A_{12}	A_{13}
X_2	A_{20}	A_{21}	A_{22}	A_{23}
X_3	A_{30}	A_{31}	A_{32}	A_{33}

$$G_k = P_0 \cdot A_{k0} + \cdots + P_n \cdot A_{kn}, \tag{1}$$

then G_k is the expected gain from the kth course of action.

To test *accuracy*, strictly,

$$\left.\begin{array}{ll} A_{kj}=1 & \text{for } k=l \\ A_{kj}=0 & \text{for } k\neq j \end{array}\right\}. \tag{2}$$

For a test of *operational value* the alphas in the matrix (Table 1) should be proportional to the operational gain from each combination of course of action and outcome. Some of these numbers may be negative indicating a loss to the operation. Perhaps all the numbers will be negative indicating costs to the operation. Hence, this matrix is sometimes called a "cost-loss matrix."

It has been claimed, directly or indirectly, that the operations of the public tend to become adjusted to the climatology of each location. Prof. L. J. Savage, Dept. of Mathematics, University of Michigan, wrote to this writer, in 1960, as follows:

> "If the probability of rain tonight is 10%, a farmer should not paint his barn today if he feels quite sure that there will be days on which this probability is only 1% during the next few weeks. If, however, the farmer lives in a climate where this probability seldom falls as low as 10% during the whole summer then he might be well advised to paint the barn now."

Professor Savage's remarks suggest that the matrix (Table 1) should, in some way, measure the forecaster's ability to recognize a departure of the conditional probabilities of events from their climatic frequencies. This brings us to the third purpose of verification: *skill in forecasting*.

Skill is defined by this writer as *the ability of the forecaster to sort or classify the weather conditions in groups so that within any one group the probability of one subsequent event is sharpened above its climatic frequency.* Supporting statements to this view of skill are given by Bryan (1960) and Sanders (1963). There is one classification of the weather, however, that requires no ability or skill to make, that is the initial class of weather. If the sky is clear at time zero then the probability of a clear sky ten hours later is sharpened above the climatic frequency of clear sky. The definition of skill needs to be modified, therefore, to eliminate rewards for the simple strategy of forecasting persistence. In symbols, the forecaster's grouping should be such that

$$P_k > {}_cP_k(I)$$

where P_k is the *a posteriori* probability of the category X_k within the forecaster's grouping or classification of the situation, and ${}_cP_k(I)$ is the conditional climatic frequency of the category X_k when the initial category is I.

3. Premise

Let us suppose that the prediction problem is to select one of the mutually exclusive events $X_0, \cdots, X_n$ and let us denote the climatic frequencies of those $(n+1)$ events as ${}_cP_0, \cdots, {}_cP_n$. (The climatic frequencies hereafter are assumed to be conditional, subject to the initial category.) After the forecaster classifies a day, there will be probabilities $P_0, \cdots, P_n$ for each mutually exclusive subsequent event. It is not necessary for our purposes that the forecaster know these *a posteriori* probabilities with a high degree of validity. We can determine $P_0, \cdots, P_n$, if we wish, after he has made a set of forecasts. (The values $P_0, \cdots, P_n$ are not the *a priori* probabilities that the forecaster might quote when he issues his forecast. We are not going to test the validity of his probability estimates.) The sum of the probabilities of mutually exclusive events is unity. Thus,

$$\sum_{i=0}^{n} {}_cP_i = \sum_{i=0}^{n} P_i = 1. \tag{3}$$

If the forecaster has no skill, then, for any group of his forecasts, there will be no expected difference between the probabilities within his grouping and the climatic frequencies. We therefore stipulate that the expected or average score of an unskilled forecaster shall be equal to some low value that we shall designate as ν (for nothing). That is,

$$\sum_{i=0}^{n} {}_cP_i \cdot A_{mi} = \nu \quad \text{for all } m. \tag{4}$$

At this point the writer offers a premise. *It must be no more profitable for the unskilled forecaster to pick one event as his forecast over another event.* Otherwise, he could "play" the system to gain a better score without skill. On the other hand, if the forecaster does indeed have skill, and if in a group of selection he makes X_k his forecast, then it must also be true that his expected score will be greater than ν. That is,

$$\sum_{i=0}^{n} P_i \cdot A_{ki} > \nu \quad \text{for} \quad P_k > {}_cP_k. \tag{5}$$

If the forecast consists of *a priori* probability estimates, it must be equally useless, on the average, to the unskilled forecaster to quote one set of odds as to quote another set. Otherwise, he will have a strategy to use, such as to quote the climatological frequencies. Unless the scoring system is deliberately designed to eliminate the advantage of such an unskilled strategy, the uncertain forecaster would soon learn how to reduce the expected penalties for error.

4. Determination of scores

The problem now is to assign numbers to the elements A_{kj} (Table 1). If we can find the scores for a specific grouping then they should be the same scores for all groupings. Let us suppose that one event X_k is increased in true (not estimated) probability and that the event X_j is decreased in probability by the same amount as the event X_k is increased. Let all other probabilities remain the same. Thus,

$$\left.\begin{aligned} (P_k - {}_cP_k) &= ({}_cP_j - P_j) > 0 \\ (P_m - {}_cP_m) &= 0 \quad \text{for } m \neq j \neq k \end{aligned}\right\}. \tag{6}$$

The scoring system must make it profitable for the forecaster to choose X_k as his forecast and nothing else. In symbols, inequality (5) must hold, and, in addition

$$\sum_{i=0}^{n} P_i \cdot A_{mi} \leqslant \nu, \quad m \neq k. \tag{7}$$

From inequality (7) and Eq. (4)

$$\sum_{i=0}^{n} ({}_cP_i - P_i) \cdot A_{mi} \geqslant 0,$$

or

$$({}_cP_j - P_j) \cdot A_{mj} \geqslant (P_k - {}_cP_k) \cdot A_{mk} + \sum_i (P_i - {}_cP_i) \cdot A_{mi}, \qquad i \neq k \neq j.$$

From (6)

$$A_{mj} \geqslant A_{mk}.$$

If, now, the roles of X_j and X_k are interchanged, then

$$A_{mk} \leqslant A_{mj},$$

TABLE 2. Matrix of scores illustrating results in Eq. (8) of text.

Forecast event	Observed event X_0	X_1	X_2	X_3
X_0	A_{00}	B_0	B_0	B_0
X_1	B_1	A_{11}	B_1	B_1
X_2	B_2	B_2	A_{22}	B_2
X_3	B_3	B_3	B_3	A_{33}

which can be true only if

$$A_{mj}=A_{mk}=B_m, \quad m\neq j\neq k. \tag{8}$$

Thus, if a forecaster makes an incorrect forecast, he will earn the same penalty score whether he barely misses or misses badly (Table 2). The scores B_m can be expected to be negative or zero.

From (4) and (8),

$$B_m=\frac{\nu-A_{mm}\cdot {}_cP_m}{1-{}_cP_m} \quad \text{for all } m. \tag{9}$$

The next question is how to find the relation between the scores from row to row. There are two scoring systems that have been proposed, those of Bryan (1960) and Gringorten (1951). Either system answers the question, "What are comparable increases in the probabilities of two events (X_k,X_m) for which the scores should make it equally profitable to the forecaster to choose either event as his forecast?"

Bryan's system makes equal arithmetic increases in the probabilities of X_k and X_m yield equal merits of forecast. That is, when

$$P_k-{}_cP_k=P_m-{}_cP_m, \tag{10}$$

then it is equally profitable to choose either X_k or X_m as the forecast.

For Eq. (4) Bryan chooses $\nu=0$. His alpha-scores become

$$\left.\begin{array}{ll} A_{mm}=K\cdot(1-{}_cP_m) & \\ A_{mj}=-K\cdot {}_cP_m & \text{for } j\neq m \end{array}\right\}, \tag{11}$$

where K is an arbitrary constant, and can be equal to 1 or 100 (Table 3).

TABLE 3. Example of scores for the forecast of mutually exclusive events X_0, X_1, X_2, X_3 whose climatic frequencies are ${}_cP_0=0.05$, ${}_cP_1=0.10$, ${}_cP_2=0.10$, ${}_cP_3=0.75$.

	Bryan Observed event				Gringorten Observed event			
Forecast event	X_0	X_1	X_2	X_3	X_0	X_1	X_2	X_3
X_0	95	−5	−5	−5	20	0	0	0
X_1	−10	90	−10	−10	0	10	0	0
X_2	−10	−10	90	−10	0	0	10	0
X_3	−75	−75	−75	25	0	0	0	1.33

Gringorten's system makes the ratios of the probabilities of X_k, X_m to their respective climatic frequencies equal for equal merits of forecast. That is, when

$$\frac{P_k}{{}_cP_k}=\frac{P_m}{{}_cP_m}>1, \tag{12}$$

then it is equally profitable to choose either X_k or X_m as the forecast. In Eq. (4) this system arbitrarily makes $\nu=1$, but not zero. Also it makes $B_0=0$, and consequently Eq. (4) gives

$$A_{00}=\frac{1}{cP_0}. \tag{13}$$

But for equal merits of forecasting X_0 and X_m,

$$A_{mm}\cdot P_m+B_m\cdot(1-P_m)=A_{00}\cdot P_0=\frac{P_0}{{}_cP_0}=\frac{P_m}{{}_cP_m}.$$

Using Eq. (9) and $\nu=1$, we derive the scores

$$\left.\begin{array}{l} A_{mm}=\dfrac{1}{{}_cP_m} \\ B_m=0 \end{array}\right\} \quad \text{for all } m. \tag{14}$$

An example of scores by the Bryan and Gringorten systems is given in Table 3.

5. Choice of scoring systems

To choose between the Bryan and Gringorten systems of scoring, suppose that a forecaster is able to forecast precipitation but is not able to decide whether the precipitation will be rain or snow. In symbols, if P_r, ${}_cP_r$, P_s, ${}_cP_s$ give the probability and climatic frequency of rain and the probability and climatic frequency of snow, then

$$(P_r+P_s)>({}_cP_r+{}_cP_s)=({}_cP_r+{}_cP_s)(1+\delta). \tag{15}$$

If $P(R|R \text{ or } S)$ is the conditional probability of R given that it will either rain or snow, then Baye's theorem gives

$$P_r=(P_r+P_s)\cdot P(R|R \text{ or } S).$$

But if the forecaster has no skill to favor rain over snow, then

$$P(R|R \text{ or } S)=\frac{{}_cP_r}{{}_cP_r+{}_cP_s}.$$

Hence,

$$\left.\begin{array}{l} P_r={}_cP_r\cdot(1+\delta) \\ P_s={}_cP_s\cdot(1+\delta) \end{array}\right\}. \tag{16}$$

Thus,

$$\left.\begin{array}{l} P_r-{}_cP_r=\delta\cdot {}_cP_r \\ P_s-{}_cP_s=\delta\cdot {}_cP_s \end{array}\right\}. \tag{17}$$

But, for Bryan's system of scoring, the merit in forecasting an event is directly proportional to the difference between the probability of the event and its climatic frequency ($P_r-{}_cP_r$ or $P_s-{}_cP_s$). Hence, if ${}_cP_r>{}_cP_s$ then Eqs. (17) show that the uncertain forecaster will have a strategy that will increase the expected score. He would predict the more frequent event, which violates the premise that there should be no strategy to benefit the unskilled forecast. On the other hand, if the merit is directly proportional to the ratio ($P_r/{}_cP_r$ or $P_s/{}_cP_s$) then Eqs. (16) show that there will be no advantage in choosing snow instead of rain or vice versa. The Gringorten system remains the only system that satisfies the premise.

Once a set of N forecasts are made using the scores of Eqs. (14), their total F is adjusted to give an ultimate rating S from zero for no skill, to one for perfect skill, by the equation

$$S=\frac{F-N}{T-N}, \qquad (18)$$

where T is the total of scores for all N accurate forecasts. Because the expected no-skill score ν was made equal to one, the expected no-skill total is N.

6. Example

To illustrate the scoring system of Eqs. (14), a set of 1098 forecasts of ceiling height at Duluth, Minn., were collected from Air Force Base forecasters during the years 1960, 1961, 1962. (Climatic differences between seasons were neglected with respect to ceiling heights.) The five mutually exclusive categories and their climatic frequencies were:

X_0	<200 ft	4%
X_1	200–450 ft	18%
X_2	500–950 ft	13%
X_3	1000–9950 ft	15%
X_4	Ceiling unlimited	50%

TABLE 4. The observed climatic frequencies of ceiling height at Duluth, Minn., in five categories (see text). The frequencies are conditional, following 18 hr after the initial condition.

Initial category	Observed category X_0	X_1	X_2	X_3	X_4	Sample size
X_0	0.21	0.33	0.09	0.12	0.25	33
X_1	0.10	0.35	0.13	0.11	0.31	186
X_2	0.03	0.25	0.14	0.14	0.44	147
X_3	0.03	0.22	0.11	0.18	0.46	147
X_4	0.02	0.11	0.09	0.20	0.58	585
Total						1098

Note that the first event was rare and the last event was common. The forecasters were required to make the forecasts to verify 3, 8, 12 and 18 hr later. The conditional climatic frequencies, for the 1098 forecasts, made 18 hr ahead of deadline, are shown in Table 4. The scores for each combination of forecast and observed events are shown in Table 5.

Fig. 1 shows the results for the ratings S. The unskilled forecast of persistence or the constant forecast of one kind of event, sometimes known as the "climatology" forecast, always achieved the rating zero. Pure

TABLE 5. The Duluth scores for each combination of forecast and observed event 18 hr later.

	Initial event: X_0					Initial event: X_1				
	Observed event					Observed event				
Forecast event	X_0	X_1	X_2	X_3	X_4	X_0	X_1	X_2	X_3	X_4
X_0	4.8	0	0	0	0	10.0	0	0	0	0
X_1	0	3.0	0	0	0	0	2.8	0	0	0
X_2	0	0	11.1	0	0	0	0	7.7	0	0
X_3	0	0	0	8.3	0	0	0	0	9.1	0
X_4	0	0	0	0	4.0	0	0	0	0	3.2

	Initial event: X_2					Initial event: X_3				
	Observed event					Observed event				
Forecast event	X_0	X_1	X_2	X_3	X_4	X_0	X_1	X_2	X_3	X_4
X_0	33.3	0	0	0	0	33.3	0	0	0	0
X_1	0	4.0	0	0	0	0	4.5	0	0	0
X_2	0	0	7.1	0	0	0	0	9.1	0	0
X_3	0	0	0	7.7	0	0	0	0	5.6	0
X_4	0	0	0	0	2.2	0	0	0	0	2.2

	Initial event: X_4				
	Observed event				
Forecast event	X_0	X_1	X_2	X_3	X_4
X_0	50.0	0	0	0	0
X_1	0	9.1	0	0	0
X_2	0	0	11.1	0	0
X_3	0	0	0	5.0	0
X_4	0	0	0	0	1.7

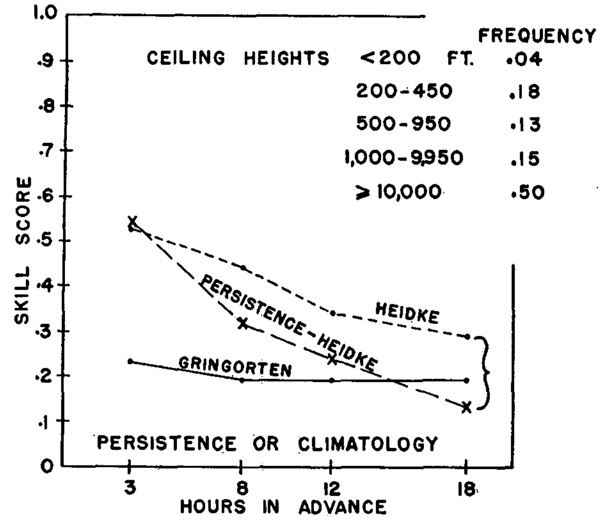

FIG. 1. Ratings (see text) for 1098 sets of Air Force forecasts of ceiling at Duluth, Minn. Solid line shows ratings by the scoring system described in this paper. Broken lines show the ratings by the scoring system of Heidke (1926).

guess work also would have earned the expected value $S=0$. But the Duluth forecasters did show skill with $S=0.19$–0.23. They used something beyond their knowledge of climatology and persistence of the weather to earn scores.

For comparison, Fig. 1 shows the skill scores by the Heidke (1926) method of verification. For the 3-hr forecasts the score is 0.53, which is no better than the score for persistence, which means that the Heidke system has failed to uncover the skill in the forecasts for 3-hr in advance. This illustrates why it is important to eliminate the advantages of unskilled strategies. By doing so we eliminate noise from the test and allow a true note of skill to emerge.

In Fig. 2 the results for the same forecasts are grouped by categories of the predictand. By the Heidke system, marked H for the forecasters' ratings and P-H for the ratings of persistence forecasts, again it would seem that the forecasters have no skill in forecasting the rare event X_0 which is denied by the other system, marked G. The forecasters are skillful in all categories.

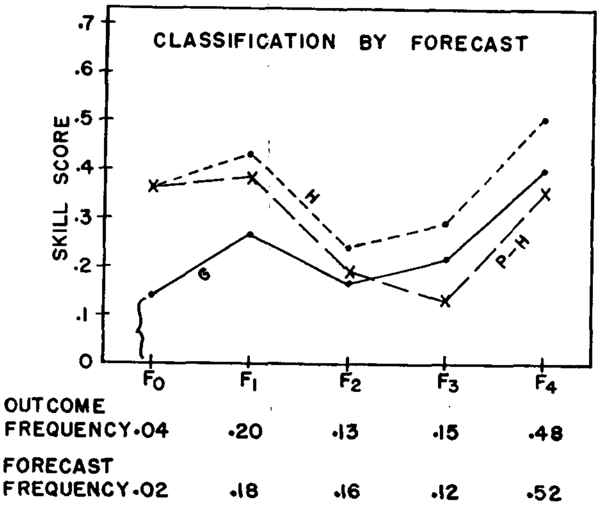

FIG. 2. Ratings for the same 1098 sets of forecasts as in Fig. 1 but classified into the five categories of ceiling.

7. General non-acceptance

The forecast system of Eqs. (14) has not enjoyed general acceptance since its publication in 1951 although its attributes have been recognized and commended occasionally. In fact, no system of verification and scoring has had universal acceptance. The reasons for the non-acceptance of this system are worth considering.

1) A system that yields skill scores of 20–30% offers a sad commentary on the state-of-the-forecasting art. The public has heard figures like 80 and 85%.

2) The scores are positive for hits, and zero for all misses. But the human instinct is to give at least a partial score for being close to accurate.

3) Individual scores are large for rare events, small for frequent events. Missing the rare forecasts can become tantalizing. More significantly, it makes the average score unstable until a large sample of forecasts is accumulated.

4) The system of scoring immediately suggests that the forecaster should have, on hand, a table of climatic frequencies. Surprisingly, not all weather stations are equipped with such climatological tables, especially tables of conditional climatology.

5) The forecaster is encouraged by the system to think in probability terms. Yet he is asked to choose only one category for each forecast and that category may not be the most likely.

6) It has been felt that forecasters should be judged by the validity of their probability estimates as well as by the sharpness of their probability estimates.

7) There is still the problem of making the forecasts operationally useful. Admittedly, the forecasts that are prepared for a verification program will not necessarily serve best in any specific operation.

The answers to these criticisms are as follows:

1) High percentages of accuracy or skill, for their own sake, are meaningless. We are concerned with relative measures, indeed to determine if there even exists a usable skill in some instances. Forecast evaluation must be based upon a large sample of predictions to yield a statistically valid result. When the statisticians's sample is large enough, then the near-hits are compensated by near-misses. The *average* score will become a truly *representative* score. And the human instinct, to give credit for partial success, will be gratified by a large sample's rewards for the statistically inevitable hits. If some individual scores are large, and some small, then it will

take simply a larger sample to produce a more stable result.

2) Operationally, the verification program's value is in the encouragement given the forecaster to work to the limit of his capability, to sharpen the probabilities of his estimates. The so-called "forecast" that is required for verification purposes need *not* be the answer that is supplied to the operations office. If, however, the administrative offices demand that the forecaster issue only one set of probabilities, the system of scoring is still applicable. If the quantities ${}_fP_0$, ${}_fP_1$, ${}_fP_2$, ${}_fP_3$ are the forecaster's *a priori* probability estimates, these estimates should be compared with the climatic frequencies. If the estimate of the event X_m is less than the climatic frequency then the forecast is interpreted as not $-X_m$. In view of the observed or verified event, these probability estimates will rate one score for each estimate. Nothing is lost by this revision of the program, although the task of verification and scoring is increased considerably.

3) With regard to *validity* of the probability estimates, is this not a secondary goal, secondary to the sharpness? Operationally speaking, two forecasters can quote differing *a priori* probability estimates. Yet the two estimates may be equally valid. As for a program that combines test of both validity and sharpness, it forces the forecaster to serve two masters. As a result, he must compromise or hedge between sharpening his estimates and making his estimates valid.

8. Concluding remark

The above emphasis on large samples to test skill is itself a weakness of the system, since it implies that we must wait a long time for results. Perhaps it will be another 10 years before a symposium similar to the January 1964 symposium is repeated, when we again shall ponder over the failure of one system to emerge as a commonly accepted system of verification. But a few items might become standardized. The tables of conditional climatology might be furnished to each forecast station and a verification program or other incentive might encourage the forecaster to consult these tables before issuing his forecasts.

REFERENCES

Gringorten, Irving I., 1951: The verification and scoring of weather forecasts. *J. Amer. Stat. Assn.*, **46**, 279–296.

Bryan, Joseph G., 1960: A proposed method for forecast evaluation. The Travelers Research Center Tech. Memo. No. 1, AF 30(635)-14459, 15 pp., preliminary, unpublished.

Heidke, P., 1926: Berechnung des Erfolges unter der Gute der Windstarkevorhersagen in Sturmwarnungsdienst. *Geografike Ann.*, **8**, 301–49.

Sanders, Frederick, 1963: On subjective probability forecasting. *J. Appl. Meteor.*, **2**, 191–201.